Abstract Title Page

Title

Assessing Teacher Effectiveness Through Dual-Rater Classroom Observations: Researchers and District Staff Partnering to Create Calibrated Performance Evaluations

Authors and Affiliations

David P. Manzeske, American Institutes for Research
Jared P. Eno, American Institutes for Research
Robert M. Stonehill, American Institutes for Research
John M. Cumming, Jefferson County (Colorado) Public Schools
Heather L. MacGillivary, Jefferson County (Colorado) Public Schools

Abstract Body

Problem/Background/Context

Federal policies (e.g., 2002 reauthorization of the Elementary and Secondary Education Act [ESEA] and the American Recovery and Reinvestment Act) posit that teacher quality is a potential leverage point for improving student achievement (U.S. Department of Education, 2010). Moreover, in the Race to the Top competition, teacher effectiveness must be based, in part, on teacher performance measured by classroom observations. This has driven many districts to adopt teacher classroom observation rubrics to meet the Race to the Top requirement.

The field has focused on validity and reliability issues related to the rubrics themselves (see, for example, Bill & Melinda Gates Foundation, 2012). However, not as much is known with respect to the best practices for employing these rubrics. Without clear guidance on how to rate teachers and without proper calibration activities, scores on these rubrics can become upwardly biased, leading to an inability to distinguish among teachers at different performance levels (see, for example, Weisberg, Sexton, Mulhern, & Keeling, 2009). When a rubric is used inconsistently, teachers may not receive useful feedback, and the rubric could lack teacher buy-in, resulting in views that the evaluation does not provide credible information. This can undermine the utility of the instrument and, more broadly, the policy initiative on which use of the rubric was founded.

Purpose/Objective/Research Question/Focus of Research

In partnership with district central office staff, a study was conducted to investigate the use of a classroom observation rubric within the context of a district's pilot teacher performance evaluation system. A component of this system was that principals and teachers' peers conduct the classroom observations. These peers were accomplished former teachers who were selected by the district and hired to observe and evaluate their fellow teachers. Some of the peer raters were retired teachers; others were on special assignment and did not have a classroom assignment. None of the peer raters had previously worked with or been in the same building with the teachers being observed.

The district wanted to know whether peer raters would use the observation rubric differently compared with principals. The district also was interested in knowing whether interrater reliability differed by rater type. Finally, the district wanted to determine whether district-selected raters, on average, were more or less lenient than principals in rating other teachers.

Improvement Initiative/Intervention/Program/Practice

Many districts are struggling to implement teacher evaluation systems in a way that generates high levels of teacher buy-in. One factor could be that some evaluation systems lack consistency in the performance information that is produced. This could be especially true for a system that uses a combination of raters other than school principals to reduce the burden that is placed on principals to observe every classroom teacher multiple times per year. Therefore, it is important to identify ways in which the evaluation system is producing inconsistent information. If the inconsistency is a function of rater type, for example, then actions can be taken to better inform rater training needs and calibration efforts. With greater consistency in the performance

information produced by evaluation systems, teachers may view the information as being more credible, and, in turn, teachers will use the information in earnest to improve their effectiveness.

Setting

The study took place in a set of 20 elementary (n = 16) and middle schools (n = 4) in a large district in the western United States. The schools had among the highest levels of socioeconomic disadvantage and lowest levels of performance in the district. These schools were participating in a pilot program that included district-employed peer raters who observed and rated teachers in the 20 schools. As part of the pilot program, each teacher had at least one formal observation by a peer rater and one by the school principal (either the principal or an assistant principal). Peer raters and principals rated teachers using a rubric based on the Charlotte Danielson Framework for Teaching. At the end of the school year, after classroom observations were conducted, peer raters and principals collaboratively assigned each teacher an overall, final evaluation rating informed by the teacher's observation ratings from that school year. Another aspect of the pilot included an opportunity for performance pay based in part on observation ratings and student academic growth on the state assessment. The performance pay system used by one half of the pilot schools gave teachers a stipend if they achieved the highest year-end evaluation rating.

Population/Participants/Subjects

This study involved a nonrandom selection of 65 teachers in fall 2011 and 87 teachers in fall 2012. Observations were conducted for Grades K–8 teachers. In fall 2011, raters included all 10 district-employed peer raters and 28 principals, with at least one principal from each school; in fall 2012, raters included all nine district-employed peer raters and 31 principals, with at least one principal from each school.

Teachers were observed for one lesson by groups of two to four raters.[*] These observations resulted in 68 pairwise comparisons between raters in fall 2011 and 120 pairwise comparisons in fall 2012. Pairwise comparisons consisted mostly of peer raters compared with principals (76 percent in fall 2011 and 79 percent in fall 2012) but also included a number of comparisons of two peer raters (24 percent in fall 2011 and 12 percent in fall 2012) or two principals (none in fall 2011 and 9 percent in fall 2012). Slightly more than 50 percent of the pairwise comparisons were for observations of teachers in schools implementing performance pay (59 percent in fall 2011 and 55 percent in fall 2012).

Research Design

To examine trends among raters and differences between groups of raters, each teacher was observed simultaneously and rated independently by at least two raters. Raters could not talk to one another when recording scores. Principals observed only those teachers in their schools, but peer raters observed at multiple schools, such that there were no disjointed subsets; a path existed in the data to connect every rater to one another.

[*] Observations in fall 2011 included groups of two raters only.

The assignment of two raters providing independent ratings of the same measurement occasion was a particular strength of the study. It enabled more precise comparisons of how different raters rated the same lessons.

An alternative to this design was to compare the scores given by each rater type from independent measurement occasions; however, when such an approach is used to compare ratings, other factors could drive differences between raters. Such a design is not common in applied settings, in part because of financial constraints. However, the researchers and the district staff worked closely together to negotiate the design parameters.

Data Collection and Analysis

Observations were conducted using a 15-item rubric, with six items classified as "professional preparation" and nine items classified as "professional techniques" (see Table 1). For all but one item, the teacher could be rated ineffective, emerging, effective, or distinguished (one item in the professional techniques section could be rated only effective or ineffective). Raters also could give an N/A (not applicable) rating if there was insufficient evidence observed on which to provide a rating. The district discouraged use of the N/A rating in fall 2012.

Dual-rater observation ratings were analyzed in three ways. First, the frequency of rater agreement was examined. There are no absolute guidelines for what constitutes sufficient frequency of agreement, but an informal guideline of 80 percent was used.

Second, Cohen's kappa—an interrater reliability statistic that adjusts for the level of agreement likely to be observed by chance (Cohen, 1960)—was calculated for each item. There are no universally accepted guidelines for interpreting kappa, in part because it depends on not only rater reliability but also other factors, such as the number of response options and how response options are used across all raters (e.g., Bakeman, McArthur, Quera, & Robinson, 1997). However, comparisons with the frequently referenced guidelines by Landis and Koch (1977) are noted. According to these guidelines, reliability from .01 to .20 is slight, from .21 to .40 is fair, from .41 to .60 is moderate, from .61 to .80 is substantial, and from .81 to .99 is almost perfect.[†]

Finally, dual-rater observation data were used to calculate each rater's severity (or leniency) by implementing the many-facet Rasch measurement model (Linacre, 1989; Linacre & Wright, 2004). Rater severity is a logit score in which a larger positive severity metric indicates higher levels of severity and a larger negative severity metric indicates less severity.

Findings/Outcomes

Descriptive analysis found that for most items, the raters indicated that they had insufficient evidence to give a rating. Items with many N/A ratings included those related to assessments (items II.d and II.e), educator knowledge (I.a and I.b), and using appropriate consequences (II.i). The N/A rating was used less in 2012 than in 2011, reflecting district guidelines to avoid this option. However, in both years, the rater pairs disagreed on whether there was sufficient evidence for a rating. For four items, the raters agreed on whether there was sufficient evidence

[†] See Cicchetti (1994) for a similar but separate set of guidelines.

for a rating less than 80 percent of the time in both years. Disagreement about the sufficiency of evidence was less in 2012 than 2011 for all but one item (see Figure 1).

When the raters agreed that there was sufficient evidence for an item, the most common rating for 14 of 15 items was effective in both years. In 2012, agreement on actual ratings (those other than N/A) ranged from 68 percent to 83 percent across items, averaging 74 percent. Kappa averaged .50 across items when the raters agreed that there was sufficient evidence. These levels of interrater reliability were higher than in 2011 for most items (10 and nine out of 15 items for agreement and kappa, respectively; see Table 2).

In both years, on average, the peer raters were more likely to give an N/A rating than principals and less likely to give a distinguished rating (see Table 3). Reliability was similar for principal-peer comparisons and peer-peer comparisons when averaged across items (see Table 4). However, reliability varied by item; for instance, in 2012, the kappa statistic for peer-to-peer comparisons was higher than for principal-peer comparison for seven items and lower for the remaining eight items. These differences exceeded .10 for 11 items and exceeded .20 for four items.

The many-facet Rasch analysis revealed that in both years, there were statistically significant differences in severity among all raters, regardless of whether the rater was a peer or a principal. This suggests that severity differed enough that the raters were not interchangeable. Differences in average severity between peers and principals were not statistically significant in 2011 but were statistically significant in 2012, with peers being more severe than principals ($p < .05$; see Table 5). The 2011 to 2012 difference in rater severity between peer raters and principals may have been driven by the fact that disagreements between peer raters and principals were more concentrated in the N/A and distinguished categories in 2012 than in 2011, when the differences were spread more evenly across the rating categories. Furthermore, in both years, severity levels among principals were more varied than those of peer raters (see Table 5). In 2012, five of the eight most severe raters (those with severity scores more than one standard deviation greater than the mean) were principals. Likewise, the eight most lenient raters (those with severity scores more than one standard deviation less than the mean) were principals. There were no statistically significant differences in rater severity between principals in schools with performance pay and principals in schools without performance pay.

Conclusions

In general, this study found that the district made progress in improving the interrater agreement and reliability of its raters. However, there is still room for improvement. This is especially evident in the rater severity differences between principals and peer raters, with principals being more lenient than peer raters. This might suggest that principals want to keep their teachers happy. Alternatively, principals drew on additional information when rating what they saw—such as preexisting beliefs—rather than rating teachers based on the instructional practices observed. Moreover, principals exhibited more varied levels of severity than peer raters. This variability might be an indication that principals were using the observation rubric with less consistency than peer raters. Based on this sample, these findings suggest that some teachers may have benefited by working in a school with a less severe principal and others may have been

disadvantaged by working at a school with a more severe principal. The strength of these findings is limited because the results are based on small samples, especially when making comparisons between groups.

The district staff used these findings to identify training needs for raters and develop rater calibration activities, which are now in place.

Further research is needed to determine the extent to which the district's calibration efforts have improved the consistency of observation ratings. But not to be overlooked is that using such types of data are an important and notable early step because districts alone often do not have the resources to undertake such research activities or calibration improvement efforts. This researcher-practitioner partnership was an important step in enhancing the district's performance evaluation system and improving educator effectiveness.

Appendices

Appendix A. References

Bakeman, R., McArthur, D., Quera, V., & Robinson, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods, 2*(4), 357–370.

Bill & Melinda Gates Foundation. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Author. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Linacre, J. M. (1989). *Many-facet Rasch measurement* (2nd ed.). Chicago, IL: MESA Press.

Linacre, J. M., & Wright, B. D. (2004). Construction of measures from many-facet data. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 296–321). Maple Grove, MN: JAM Press.

U.S. Department of Education. (2010). *A blueprint for reform: The reauthorization of the Elementary and Secondary Education Act*. Washington, DC: Author. Retrieved from http://www2.ed.gov/policy/elsec/leg/blueprint/blueprint.pdf

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on teacher effectiveness.* New York, NY: TNTP.

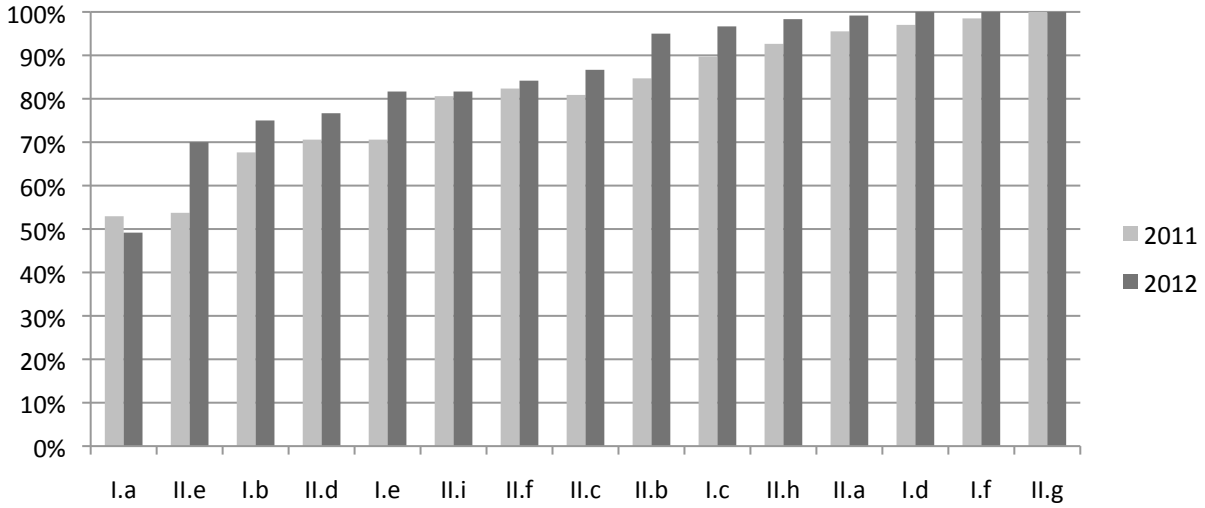
Appendix B. Tables and Figures

Table 1. Observation Rubric Items

Section	Item Number	Item Description
Professional Preparation	I.a	Demonstrates accurate, up-to-date, and extensive knowledge of subject(s).
	I.b	Demonstrates knowledge of how subject matter/disciplines are integrated.
	I.c	Implements research-based best practices.
	I.d	Develops lesson plans incorporating effective lesson design.
	I.e	Implements district-adopted curriculum through alignment of resources and assessments.
	I.f	Maximizes available instructional time.
Professional Techniques	II.a	Communicates to students expectations for learning.
	II.b	Models and facilitates higher level thinking, problem solving, creativity, and flexibility.
	II.c	Adapts instruction to meet the instructional needs of all students.
	II.d	Administers all building, district, and state assessments with fidelity.*
	II.e	Uses a variety of assessments to make instructional decisions.
	II.f	Explicitly communicates criteria for student success.
	II.g	Develops a safe and welcoming learning environment.
	II.h	Collaboratively develops, models, and communicates clear expectations for student behavior within a learning environment.
	II.i	Develops and carries out appropriate consequences in the classroom.

*A rating of only effective or ineffective was possible on this item.

Figure 1. Rater Agreement on Sufficient Evidence for 2011 and 2012



Note. In 2011, the sample sizes were $n = 67 \pm 1$ observations, except for II.d, for which $n = 59$. In 2012, the sample size was $N = 120$.

Table 2. Interrater Reliability, Excluding N/A Category

	Agreement (no N/A)			Kappa (no N/A)			*N*	
Item	**2011**	**2012**	**Difference**	**2011**	**2012**	**Difference**	**2011**	**2012**
I.a	81%	83%	+2%	.60	.67	+.07	16	41
I.b	49%	71%	+22%	.25	.51	+.27	33	78
I.c	73%	74%	+1%	.57	.57	.00	59	116
I.d	75%	74%	+0%	.56	.56	.00	55	120
I.e	74%	78%	+4%	.46	.48	+.02	31	91
I.f	75%	68%	-6%	.63	.51	-.13	67	120
II.a	73%	74%	+1%	.54	.53	-.01	64	119
II.b	69%	74%	+5%	.53	.56	+.03	64	113
II.c	67%	72%	+6%	.48	.56	+.08	60	98
II.e	50%	73%	+23%	.11	.21	+.09	8	22
II.f	72%	78%	+6%	.55	.62	+.08	50	89
II.g	85%	77%	-8%	.73	.54	-.19	66	120
II.h	78%	73%	-5%	.58	.40	-.18	54	115
II.i	77%	72%	-5%	.63	.32	-.32	39	85

Note. Item II.d is not shown because the number of evaluations that were not N/A was too small to calculate reliability statistics for this item.

Table 3. Average Frequency of Ratings for Principal and Peer Raters

Average Score Distribution	N/A	Ineffective	Emerging	Effective	Distinguished
Principals 2011	17%	5%	27%	42%	10%
Peers 2011	29%	4%	23%	38%	6%
Difference	*-12%*	*1%*	*4%*	*4%*	*4%*
Principals 2012	16%	2%	23%	47%	12%
Peers 2012	22%	3%	24%	46%	5%
Difference	*-6%*	*-1%*	*-1%*	*1%*	*7%*

Table 4. Average Kappa and Sample Size for Rater Combinations for 2011 and 2012

	Average Kappa and *N*	
Rater combination	2011	2012
Principal/peer	.39 (52[a])	.38 (95)
Peer/peer	.40 (16[b])	.39 (14)

Note. The *N* for each comparison is shown in parentheses.
[a]For some items, *N* varied, from 45 to 52, with the mode being 52.
[b]For some items, *N* varied, from 14 to 16, with the mode being 16.

Table 5. Average Rater Severity by Rater Type

Role Type	*N*	Average Severity[a]	Standard Deviation	Least Severe	Most Severe
2011					
Peer	10	.06	.75	-1.34	1.38
Principal	28	-.02	1.22	-3.55	2.34
Total	38	.08 (diff)	1.11		
2012					
Peer	9	.45	.47	-.36	1.10
Principal	31	-.13	.77	-1.38	1.48
Total	40	.59 (diff)	.72		

[a]A larger positive severity metric indicates higher levels of severity; a larger negative severity metric indicates less severity.